



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
09/837,158	04/19/2001	Karen Mae Holland	ARC000018US1	8446
48146	7590	01/04/2006	EXAMINER	
MCGINN INTELLECTUAL PROPERTY LAW GROUP, PLLC 8321 OLD COURTHOUSE ROAD SUITE 200 VIENNA, VA 22182-3817			BLACKWELL, JAMES H	
			ART UNIT	PAPER NUMBER
			2176	

DATE MAILED: 01/04/2006

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Application No.	Applicant(s)	
	09/837,158	HOLLAND ET AL.	
	Examiner	Art Unit	
	James H. Blackwell	2176	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --
Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

- 1) Responsive to communication(s) filed on 11 October 2005.
 2a) This action is FINAL. 2b) This action is non-final.
 3) Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

- 4) Claim(s) 1-12 and 14-24 is/are pending in the application.
 4a) Of the above claim(s) 13 is/are withdrawn from consideration.
 5) Claim(s) _____ is/are allowed.
 6) Claim(s) 1-12 and 14-24 is/are rejected.
 7) Claim(s) _____ is/are objected to.
 8) Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

- 9) The specification is objected to by the Examiner.
 10) The drawing(s) filed on 09 April 2001 is/are: a) accepted or b) objected to by the Examiner.
 Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).
 Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
 11) The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

- 12) Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
 a) All b) Some * c) None of:
 1. Certified copies of the priority documents have been received.
 2. Certified copies of the priority documents have been received in Application No. _____.
 3. Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

- 1) Notice of References Cited (PTO-892)
 2) Notice of Draftsperson's Patent Drawing Review (PTO-948)
 3) Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08)
 Paper No(s)/Mail Date _____.
 4) Interview Summary (PTO-413)
 Paper No(s)/Mail Date. _____.
 5) Notice of Informal Patent Application (PTO-152)
 6) Other: _____.

DETAILED ACTION

1. This Office Action is in response to Amendment received 10/11/2005 with an original priority date of **04/19/2001**.
2. Claims 1-12, and 14-24 are pending. Claims 1, 17, and 23 are independent claims.
3. Claim 24 has been added by the Applicant.
4. The objection to claims 12, 14, and 21 has been addressed in part, by converting Claim 14 to independent form. However, in lieu of a general review of the Clustering art, and an update of the prior art search, such objections have been withdrawn. These claims have now been rejected.
5. The rejection of Claims 1-12, and 14-16 under 35 U.S.C. 101 have been withdrawn.

Claim Rejections - 35 USC § 103

6. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negated by the manner in which the invention was made.

7. Claims 1-12 and 14-24 are rejected under 35 U.S.C. 103(a) as being unpatentable over Lewak et al. (hereinafter Lewak, U.S. Patent No. 5,544,360 filed 02/03/1995) in view of Goldszmidt et al. (hereinafter Goldszmidt, "A Probabilistic

Approach to Full-Text Document Clustering", 1998, Technical Report ITAD-433-MS-98-044, SRI International).

In regard to independent Claim 1 (and similarly independent Claims 17, and 23), Lewak teaches generating a dictionary of keywords in said text documents in that the user, or an automated process (Col. 9, lines 50-55) analyzes each uncategorized file and can define categories (keywords) from those documents (Col. 8, lines 14-15; 61-65).

It is noted that the files taught by Lewak can contain files containing text along with other types of files (assuming that Fig. 1 represents files typical of those contemplated by the invention).

It is also noted that the categories (keywords) contained in a list that the user or a automated system that can be assigned can contain the *structured variables* as claimed (see Fig. 5, items listed in box containing item 52, specifically categories containing dates).

Lewak also teaches forming categories of said text documents using said dictionary and an automated algorithm in that the user, or an automated process can further group files containing similar sub-groupings together (Col. 9, lines 56-67; Col. 10, lines 1-10).

Art Unit: 2176

Lewak also teaches *counting occurrences of said structured variables, said categories, and combinations of said structured variables and said categories for said text documents* in that each category has an associated data structure record, which, among other things, stores how many files use that category along with linking and identifiers of the categories assigned to (Col. 5, lines 40-60). Given that such data is kept implies that such occurrences were tabulated.

Lewak fails to teach *calculating probabilities of occurrences of said combinations of structured variables and categories to identify a relationship between said text documents and said structured variables*. However, Goldszmidt teaches a similarity measure based on probability (an overlap measure), which measures the degree of overlap between pairs of documents (p. 4, Sec. 2, Equation 1). It would have been obvious to one of ordinary skill in the art at the time of invention to combine the teachings of Lewak and Goldszmidt as both documents discuss aspects of grouping similar documents together. Adding Goldszmidt provides the benefit of computing probabilities used to measure the similarity between text-containing files (documents) to determine the similarity between them providing a gauge of how well the chosen categories (including structured variables) define the document content.

In regard to dependent Claim 2, Lewak teaches that said algorithm comprises a keyword occurrence algorithm and wherein each of said categories comprises a category of text documents in which a particular keyword occurs (Col. 8, lines 61-67; Col. 9, lines 1-4; 50-67; Col. 10, lines 1-10; categories (keywords) are further grouped based on sub-groupings, each sub-grouping containing similar documents, based on their categories (keywords))

In regard to dependent Claim 3, Lewak fails to teach that said algorithm comprises a clustering algorithm and wherein each of said categories comprises a category of said text documents containing a particular cluster. However, Goldszmidt teaches both hierarchical agglomerative clustering as well as iterative clustering (such as K-means)(Pgs. 10-11, Sec. 3, 3.1, 3.2). It would have been obvious to one of ordinary skill in the art at the time of invention to combine the teachings of Lewak and Goldszmidt as both documents discuss aspects of grouping similar documents together. Adding Goldszmidt provides the benefit of using well known clustering techniques to group categories together and to compute probabilities used to measure the similarity between text-containing files (documents) to determine the similarity between them providing a gauge of how well the chosen categories (including structured variables) define the document content.

In regard to dependent Claim 4, Lewak fails to teach that *said clustering algorithm comprises a k means algorithm*. However, Goldszmidt teaches both hierarchical agglomerative clustering as well as iterative clustering (such as K-means)(Pgs. 10-11, Sec. 3, 3.1, 3.2). It would have been obvious to one of ordinary skill in the art at the time of invention to combine the teachings of Lewak and Goldszmidt as both documents discuss aspects of grouping similar documents together. Adding Goldszmidt provides the benefit of using well known clustering techniques to group categories together and to compute probabilities used to measure the similarity between text-containing files (documents) to determine the similarity between them providing a gauge of how well the chosen categories (including structured variables) define the document content.

In regard to dependent Claim 5, Lewak teaches *said forming said categories comprises inputting a predetermined number of categories* (Col. 5, lines 28-31).

In regard to dependent Claim 6, Lewak fails to teach that *said forming said categories comprises: generating a sparse matrix array containing a count of each of said keywords in each of said text documents*. However, generating a sparse matrix in this way is well known in the art and is typically a crucial part of most clustering algorithms.

In regard to dependent Claim 7, Lewak teaches that *said keywords comprise at least one of words and or phrases, which occur a predetermined number of times in, said text documents* (see Figs. 3-5 categories (keywords) can be words or phrases).

In regard to dependent Claim 8, Lewak fails to teach *said calculating probabilities comprises using a Chi squared function*. However, Goldszmidt teaches using a Chi-Squared test as part of the analysis of clustering methods (p. 15, 3rd paragraph). It would have been obvious to one of ordinary skill in the art at the time of invention to combine the teachings of Lewak and Goldszmidt as both documents discuss aspects of clustering techniques. Adding Goldszmidt provides the benefit of using statistical measures to analyze clustering results.

In regard to dependent Claim 9, though Lewak fails to specifically teach that *said generating a dictionary of keywords comprises: first parsing text in said text document to identify and count occurrences of words; storing a predetermined number of frequently occurring words; second parsing text in said text documents to identify and count occurrences of phrases; and storing a predetermined number of frequently occurring phrases*, Lewak does either manually or automatically perform and provide a mechanism for compiling such a dictionary that involves viewing/parsing each of the uncategorized documents and determining, based on the subject matter (to include contemplating the number and meaning of descriptive term(s) or groups of terms) whether or not the term(s) or groups of terms are significant to describing the text document content. Thus, one of ordinary skill in the art at the time of invention would have considered such a method of compiling a list of keywords to be obvious based on well known and widely used techniques such as is contemplated by Lewak and as claimed.

In regard to dependent Claim 10, Lewak fails to teach that *said frequently occurring words and phrases are stored in a hash table*. However, it is typical to use hash tables as data structures, especially when the storage of vectors and matrices involved with clustering algorithms to enable their efficient storage and subsequent evaluation on a computer.

In regard to dependent Claim 11, Claim 11 contains subject matter that is similar to that found in Claims 1 (and similarly Claims 17 and 23) and 6, and is rejected along similar lines of reasoning.

In regard to dependent Claim 12, Lewak fails to teach that *said relationships comprise said combinations of structured variables and categories having a lowest probability of occurrence*. However, it is notoriously well known in the art that measures of whether or not two objects are grouped together or not depend on how closely or how distant characteristics of two objects are in comparison to one another. Those that are distant in terms of their similarities would translate to having a low probability of occurrence. Likewise, such similarity measures would also allow one to deduce how likely the clustering of two objects is the result of randomness.

In regard to independent Claim 14, Claim 14 reflects the method for identifying relationships between text documents and structured variables pertaining to said text documents as claimed in Claim 1 (and similarly Claims 17, and 23) and Claim 12, and is rejected along the same rationale.

In regard to dependent Claim 15 (and similarly dependent Claim 19), and dependent Claim 16 (and similarly dependent Claim 20), Lewak teaches that said *structured variables comprise predetermined time intervals and said predetermined time intervals comprise one of days, weeks, months and years* (see Figs. 3-5, category phrases involving time, date 10,92; dataq 6 oct; dataquick Oct. 9.92).

In regard to dependent Claim 18, Lewak does not specifically teach a *memory for storing occurrences of said structured variables, categories and structured variable/category combinations and probabilities of occurrences of said structured variable/category combinations.* However, it would have been obvious to one of ordinary skill in the art at the time of invention to assume that such data would have to have been stored on some media such as memory, disk, or other computer storage, providing the benefit of ready access to the data for processing on a computer.

In regard to dependent Claim 21, Claim 21 contains subject matter similar to that found in Claim 14, and is rejected for similar reasons.

In regard to dependent Claim 22, Lewak teaches that *said relationships comprise statistically significant relationships*. However, Goldszmidt teaches both hierarchical agglomerative clustering as well as iterative clustering (such as K-means)(Pgs. 10-11, Sec. 3, 3.1, 3.2). The determination of similarity is at the heart of most clustering algorithms because it is that measure that allows those algorithms to group similar documents together. Even if done manually, as in the teaching of Lewak, a human being of ordinary skill would have been able to produce groupings of documents that would have been statistically significant. It would have been obvious to one of ordinary skill in the art at the time of invention to combine the teachings of Lewak and Goldszmidt as both documents discuss aspects of grouping similar documents together. Adding Goldszmidt provides the benefit of using well known clustering techniques to group categories together and to compute probabilities used to measure the similarity between text-containing files (documents) to determine the similarity between them providing a gauge of how well the chosen categories (including structured variables) define the document content.

In regard to dependent Claim 24, Lewak teaches that *said structured variables comprise structured data* see Figs. 3-5, category phrases involving time, date 10,92; dataq 6 oct; dataquick Oct. 9.92).

Response to Arguments

8. Applicant's arguments with respect to claims 1-24 have been considered but are moot in view of the new ground(s) of rejection. Respectfully, the arguments made by the Applicant are substantially based on the validity of the combination of the prior art of Allan and Goldszmidt. Applicant argues that the combination of the prior art of Allan and Goldszmidt are not combinable as they both relate clustering different types of content. The Examiner respectfully agrees and withdraws the rejection. However, the Examiner now introduces the prior art of Lezak in combination with Goldszmidt. Both of these references relate to document grouping (clustering). This combination also covers the limitation of *the calculation of probabilities of occurrence of the combinations of structured variables and categories to identify a relationship between the text documents and the structured variables.*

Conclusion

9. Any inquiry concerning this communication or earlier communications from the examiner should be directed to James H. Blackwell whose telephone number is 571-272-4089. The examiner can normally be reached on Mon-Fri.

10. If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Heather R. Herndon can be reached on 571-272-4136. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

11. Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

James H. Blackwell
12/20/2005

William F. Bashore
WILLIAM BASHORE
PRIMARY EXAMINER
12/22/2005